KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY

Bc. Josef Baloun

Diplomová práce

Inženýrská informatika
Medicínská informatika
2019/2020

Vedoucí práce:
doc. Ing. Pavel Král, Ph.D

Segmentace stran rukopisných dokumentů

Abstrakt

Analýza stran dokumentů hraje významnou roli v procesu jejich elektronického zpřístupnění. Dokonce i v současné době může představovat nelehkou výzvu pro historické ručně psané dokumenty vzhledem k jejich různorodé struktuře a možné degradaci kvality. V rámci této práce je vypracován přehled možných metod pro řešení tohoto problému a vytvořena datová sada složená ze stran ručně psaných kronik. Dále je navržen prototyp systému pro analýzu stran dokumentů. Segmentace a klasifikace do tříd text, obrázek a pozadí jsou řešeny označením každého obrazového bodu strany dokumentu vhodnou třídou. Základem prototypu je plně konvoluční neuronová síť založená na síti U-Net. Nejlepších výsledků bylo dosaženo s prototypem, pro který bylo nastaveno zpracování celých stran dokumentů, bylo provedeno váhování chybové funkce a byla automaticky rozšířena trénovací množina.

Úvod

V současné době je patrná značná snaha o digitalizaci a elektronické zpřístupnění dokumentů ve většině oblastí. Této snahy se týká i projekt *Bavorsko-česká síť digitálních historických pramenů*, jehož cílem je za pomoci rozsáhlé digitalizace a webové prezentace spojit do jednoho virtuálního celku v minulosti násilně roztržené archiválie státních archivů České republiky a Bavorska.

Základním krokem k elektronickému zpřístupnění je naskenování stran dokumentů. Dalším krokem může být analýza stran dokumentů (angl. *page layout analysis*), která se skládá ze segmentace strany dokumentu na homogenní komponenty a jejich klasifikace např. na bloky textu a obrázky. Dále může být na textové bloky aplikováno optické rozpoznávání znaků (*OCR*) a jejich převod do strojově čitelné podoby. Původní stránka dokumentu tak bude převedena do značkované, strojově čitelné podoby umožňující efektivní vyhledávání v jejím obsahu. Je snadné si představit, kolik práce by to ušetřilo historikům. V současné době je pro moderní tištěné dokumenty tento proces na velice dobré úrovni, ale problém nastává u starších a ručně psaných dokumentů vzhledem k jejich různorodé struktuře a degradované kvalitě.

Tato práce se týká zmiňované analýzy stran ručně psaných dokumentů, které mohou vhodně reprezentovat např. kroniky z 19. století.

Východiska, analytická část

Segmentací stran se rozumí úloha extrahování homogenních komponent z obrázku stránky dokumentu. Homogenní komponenty mohou představovat např. textové bloky, řádky textu, tabulky a obrázky. Úloha segmentace stran nezahrnuje klasifikaci komponent, ale je důležité pochopit, že tyto úlohy nelze oddělit.

Prostudovaná řešení zahrnují metody založené např. na spojených komponentách. Dále je pro segmentaci možné využít skeletizace pozadí. Zde se předpokládá snadné rozdělení pozadí od popředí. Další metody zahrnují extrakci příznaků a jejich použití pro klasifikaci. Nejúspěšnější metody řešení tohoto problému využívají plně konvoluční neuronové sítě.

Hlavní aspekty realizace

Základem systému pro segmentaci stran rukopisných dokumentů je plně konvoluční neuronová síť, která je založená na síti *U-Net*.

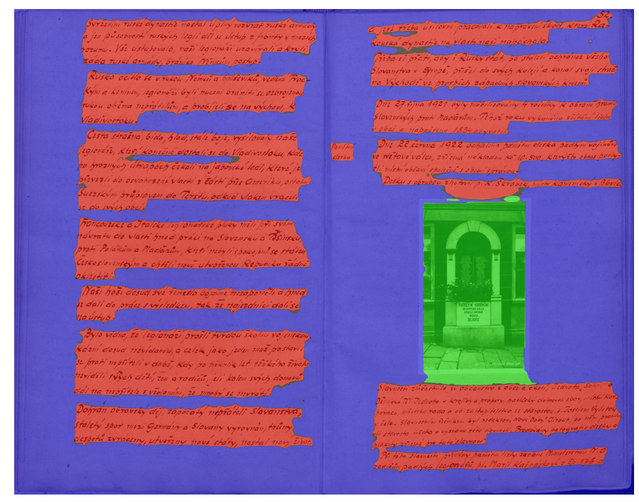
K problému segmentace je přistupováno jako k *pixel-labeling* problému, kde je každý pixel zařazen do odpovídající třídy.

Pro úspěšné natrénování sítě je nutná odpovídající datová sada, která ale nebyla nalezena. Z tohoto důvodu byla vytvořena datová sada založená na stranách kronik poskytnutých z portálu Porta fontium.

Postup zpracování začíná načtením vstupního obrázku ve stupních šedi a úpravou jeho rozměrů pro splnění požadavků na vstup, které jsou dány architekturou sítě. Následuje predikce segmentace pomocí natrénované sítě. Výstup sítě tvoří hodnoty poslední vrstvy s aktivační funkcí sigmoid. Tím se pro každou požadovanou třídu získá maska se stejnými rozměry jako rozměry vstupu sítě. Tyto masky se nejdříve zvětší na původní rozměry obrázku a následně jsou prahovány. Výsledkem segmentace jsou binární masky pro jednotlivé třídy, jejichž rozměry odpovídají původnímu obrázku.

Dosažené výsledky

Prototyp systému pro segmentaci dosáhl výborných výsledků nejen na stranách kronik. Na testovací sadě bylo dosaženo skvělých 0,908 IoU a 0,991 FgPA.



Ukázka segmentace strany kroniky: červeně text, zeleně obrázek a modře pozadí

Závěr

Velkou výhodou implementovaného řešení je možnost segmentace prakticky libovolných rozměrů strany.

Architektura sítě se díky zpracování celých stran a využití paddingu v konvolučních vrstvách dokáže velmi dobře vypořádat se šumem na okraji obrázků stran.

Zajímavou vlastností sítě je i schopnost generalizace. Přestože je síť trénovaná na starších ručně psaných kronikách, dokáže dobře segmentovat i moderní tištěné strany. Dobrých výsledků je dosaženo dokonce i se sítí natrénovanou na pouhých šesti stranách kronik.

FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI